Supervised Learning Algorithm

Classification : K-nearest neighbors algorithm(KNN)

BY Gayatri Jondhale

Contents

- What is KNN?
- Why do we need KNN?
- How do we choose the factor ‘k’?
- When do we use KNN?
- How does KNN algorithm work?
- Examples
- Advantages & Disadvantages
- Application

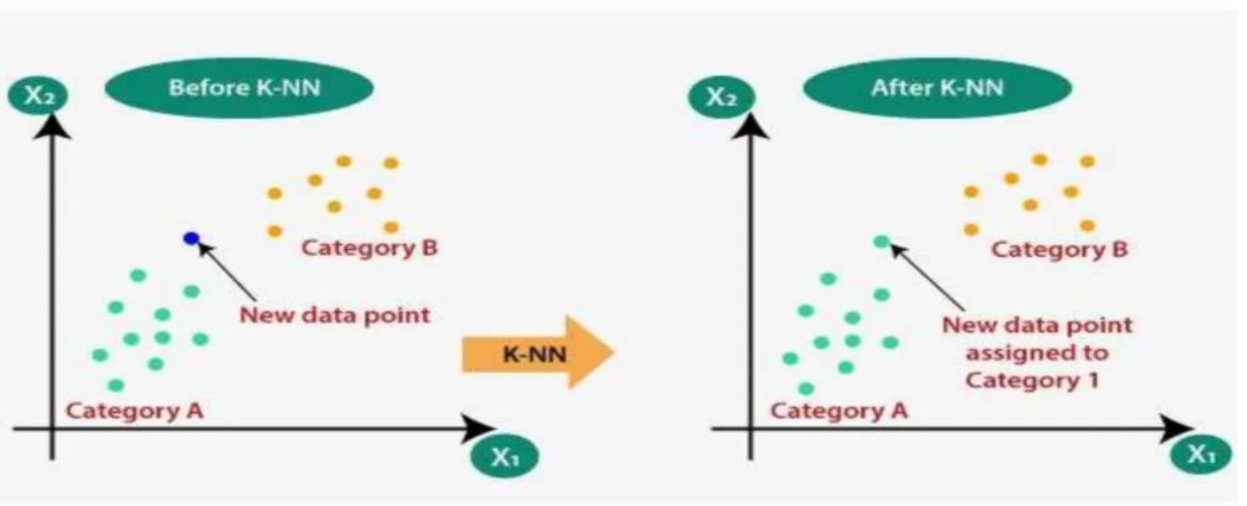
What is KNN Algorithm?

The KNN algorithm classifies a data point by finding its nearest neighbors and assigning it to the majority class of those neighbors.

KNN algorithm assume the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories .KNN is non parametric learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification , it performs an action on the dataset.

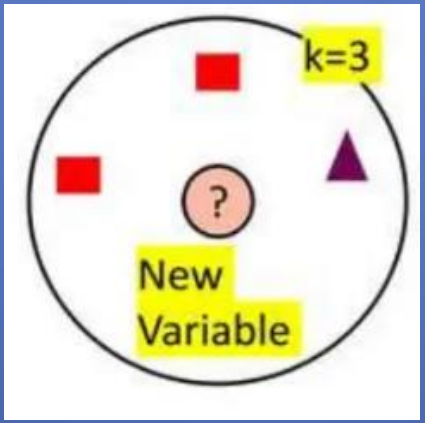
Why do we need a KNN Algorithm?

Suppose there are two categories. Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this types of problem, we need a KNN algorithm. With the help of KNN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



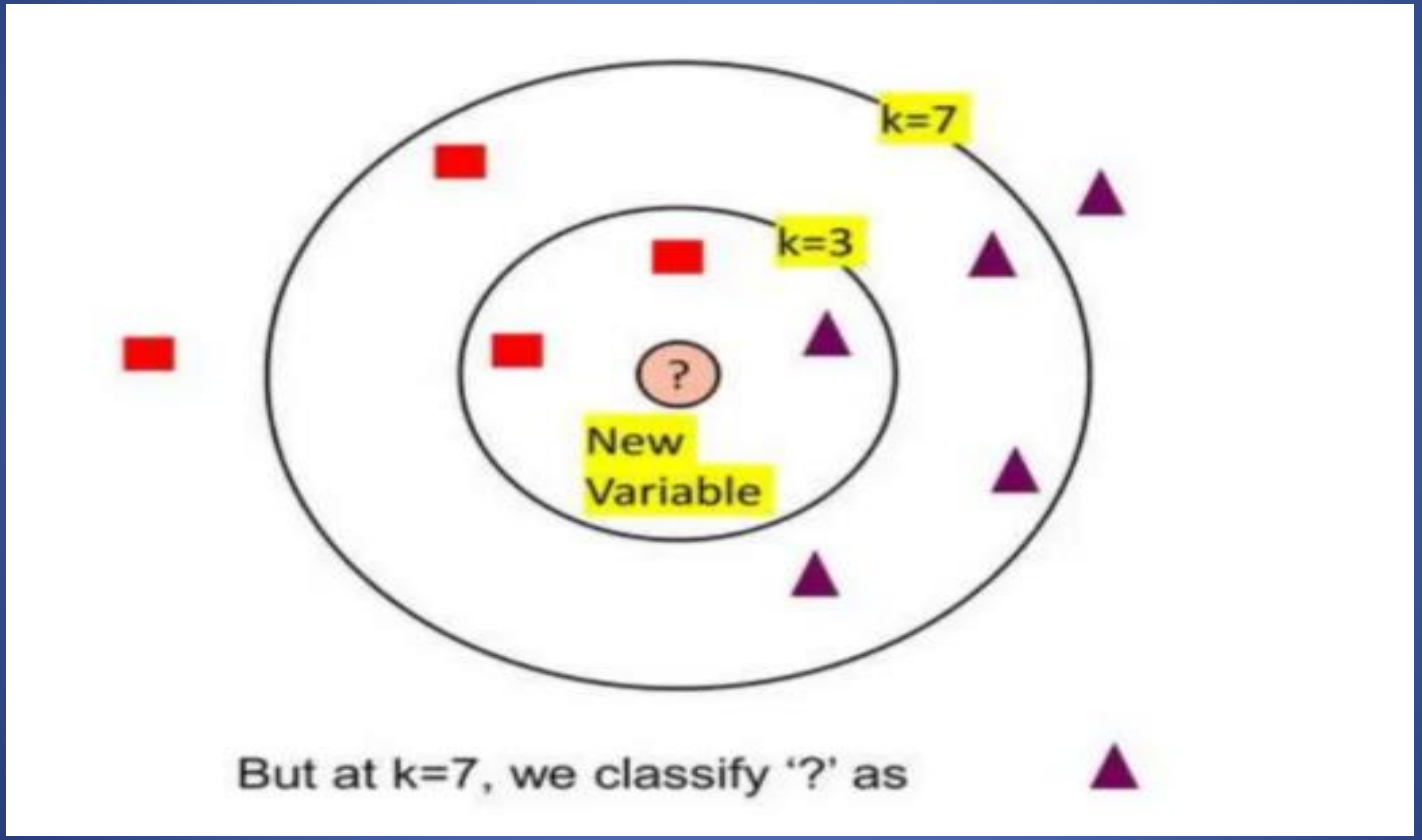
How do we choose the factor ‘k’?

KNN Algorithm based on feature similarity: choosing the right value of k is a process called parameter tuning and is important for better accuracy.



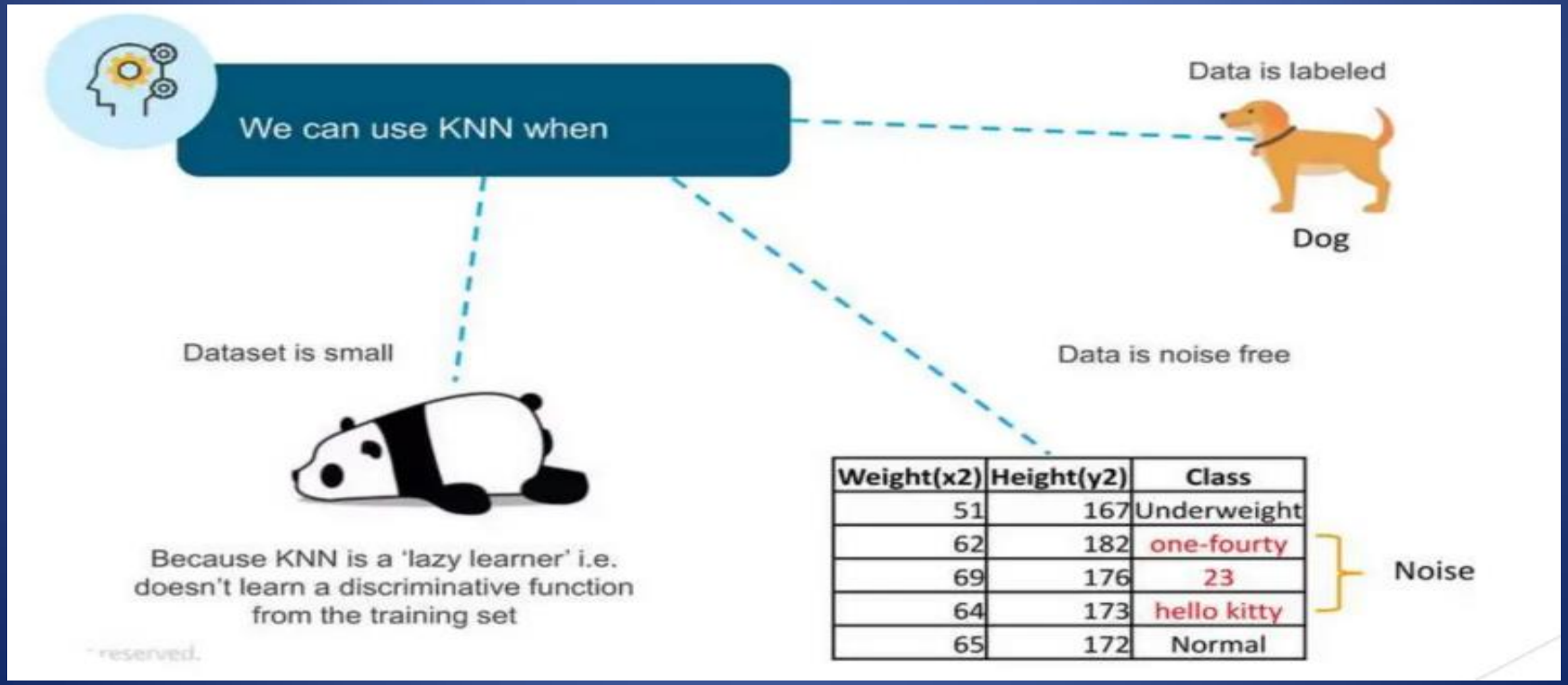
So at k=3 we can classify ‘?’ as ■

How do we choose the factor ‘k’?



When Do we use KNN Algorithm?

KNN can be used for both classification and regression predictive problems . However, it is more widely used in classification problems in the industry.

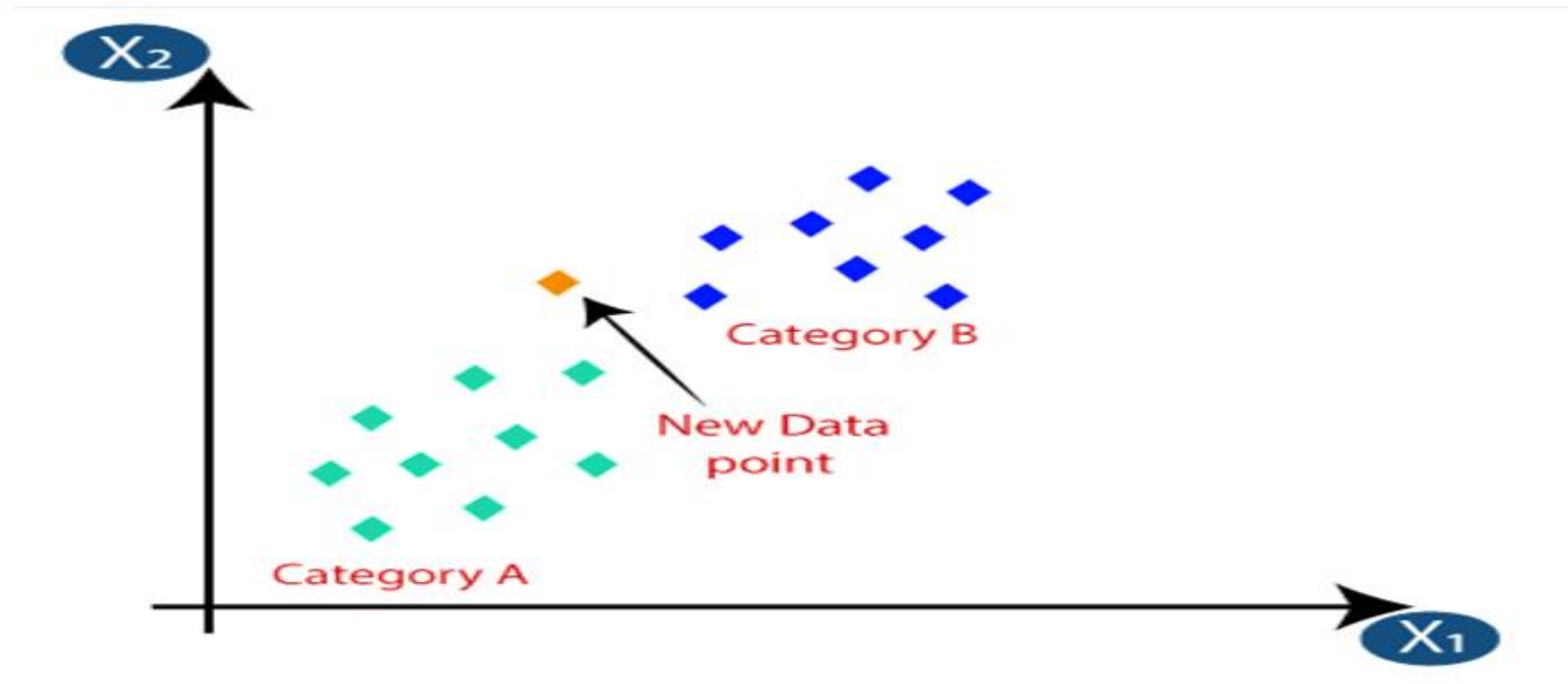


How does K-NN work?

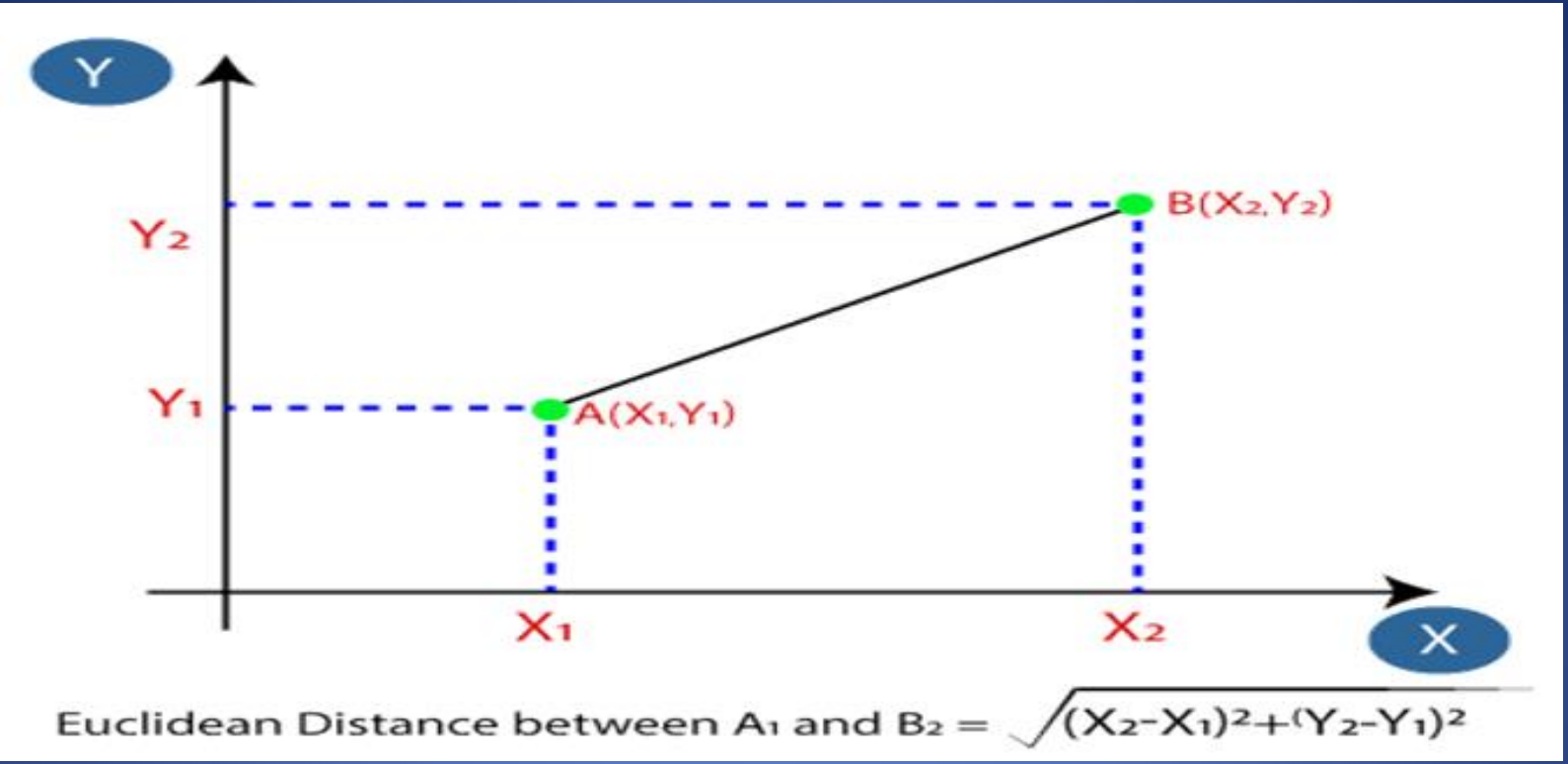
The KNN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

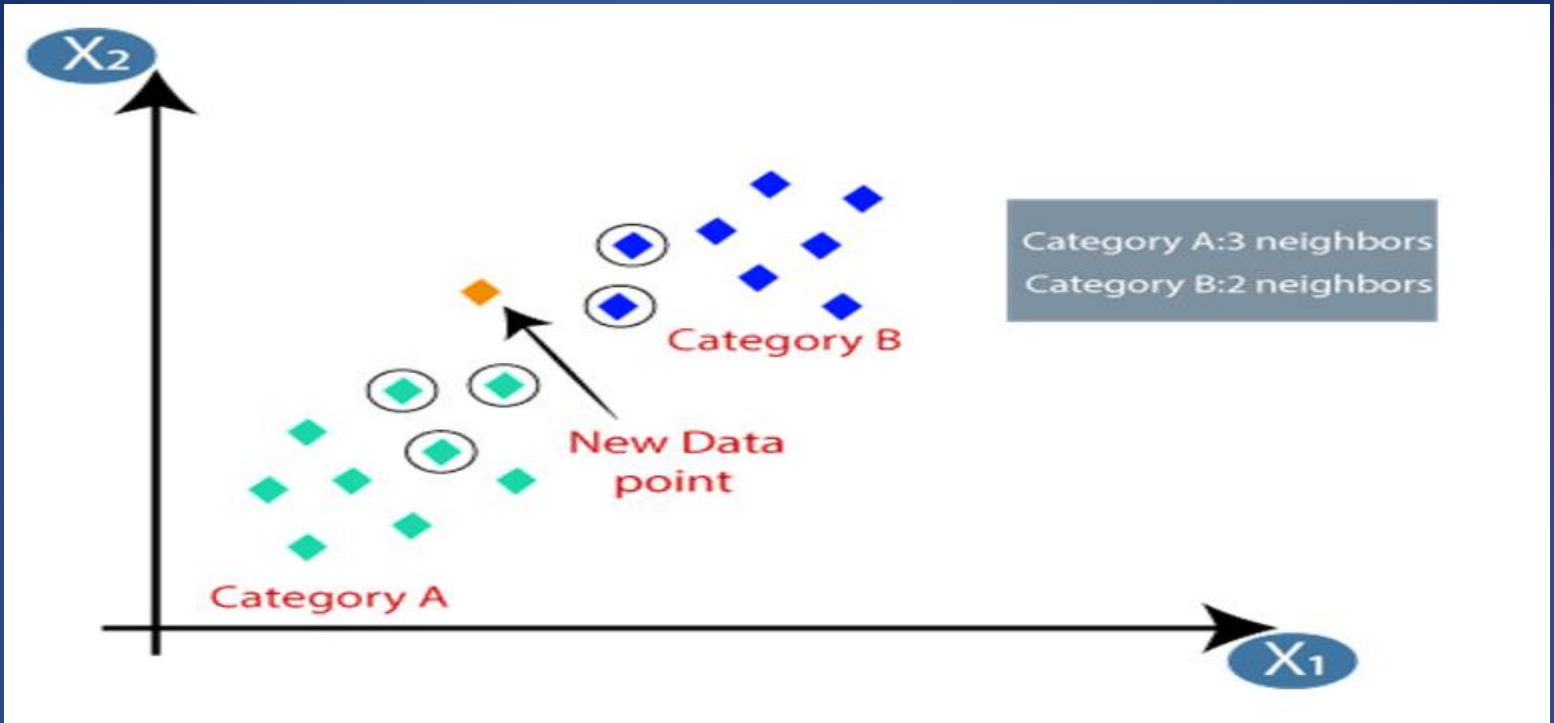
Suppose we have a new data point and we need to put it in the required category. Consider the below image:



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Example:

Name	Acid Durability	Strength	Class	Distance	Rank
Type-1	7	7	Bad	4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2
Test Data	3	7	?		

The Distance Formula:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$d1=\sqrt{(7-3)^2+(7-7)^2}=4$$

$$d2=\sqrt{(7-3)^2+(4-7)^2}=5$$

$$d3=\sqrt{(3-3)^2+(4-7)^2}=3$$

$$d4=\sqrt{(1-3)^2+(4-7)^2}=3.6$$

Based on two neighbors , Good

Example :

Height(cm)	Weight(KG)	Class	Distance	Rank
169	58	Normal	1.4	1
170	55	Normal	2	2
173	57	Normal	3	3
174	56	Underweight	4.1	4
167	51	Underweight	6.7	5
173	64	Normal	7.6	6
170	57	?		

The Distance Formula:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$d1=\sqrt{(170-169)^2+(57-51)^2}=6.8$$
$$d2=\sqrt{(170-170)^2+(57-55)^2}=2$$
$$d3=\sqrt{(170-173)^2+(57-57)^2}=3$$
$$d4=\sqrt{(170-174)^2+(57-57)^2}=4$$

now lets calculated the nearest neighbor at K=3

170 cm 57 kg = Normal

Advantages of KNN Algorithm:

1. It is simple to implement.
2. It is robust to noisy training data
3. It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

1. Always need to determine the value of K which may be complex some time.
2. The computation cost is high because of calculating between the data point for all the training samples.

Application of KNN:

The following are some of the areas which KNN can be applied successfully-

Banking System:

KNN can used in banking system to predict weather an individual is fit for loan approval? Does that individual have the characteristics similar to the defaulters one or not.

Calculating Credit Ratings:

KNN algorithms an be used to find an individual's credit rating by comparing with the persons having similar traits